Hidden Treasures: Opening Data in PhD Dissertations in Social Sciences and Humanities

Hélène Prost, Cécile Malleret, Joachim Schöpfel

JLSC

ISSN 2162-3309 | JLSC is published by the Pacific University Libraries | http://jlsc-pub.org

Volume 3, Issue 2 (2015)

Hidden Treasures: Opening Data in PhD Dissertations in Social Sciences and Humanities

Hélène Prost, Cécile Malleret, Joachim Schöpfel

Prost, H., Malleret, C., & Schöpfel, J. (2015). Hidden Treasures: Opening Data in PhD Dissertations in Social Sciences and Humanities. *Journal of Librarianship and Scholarly Communication,* 3(2), eP1230. http://dx.doi.org/10.7710/2162-3309.1230

External Data or Supplements:

[Forthcoming] Included in: "GreyNet Enhanced Publications Project", EASY, https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:51624

ISSN 2162-3309 10.7710/2162-3309.1230

RESEARCH

Hidden Treasures: Opening Data in PhD Dissertations in Social Sciences and Humanities

Hélène Prost
Information Professional, INIST (CNRS)
Cécile Malleret
Academic Librarian, University of Lille 3
Joachim Schöpfel
Senior Lecturer, University of Lille 3

PURPOSE The paper provides empirical evidence on research data submitted together with PhD dissertations in social sciences and humanities. **APPROACH** We conducted a survey on nearly 300 print and electronic dissertations in social sciences and humanities from the University of Lille 3 (France), submitted between 1987 and 2013. **FINDINGS** After a short overview on open access to electronic dissertations, on small data in dissertations, on data management and curation, and on the challenge for academic libraries, the paper presents the results of the survey. Special attention is paid to the size of the research data in appendices, to their presentation and link to the text, to their sources and typology, and to their potential for further research. Methodological shortfalls of the study are discussed, and barriers to open data (metadata, structure, format) and legal questions (privacy, third-party rights) are addressed. The conclusion provides some recommendations for the assistance and advice to PhD students in managing and depositing their research data. **PRACTICAL IMPLICATIONS** Our survey can be helpful for academic libraries to develop assistance and advice for PhD students in managing their research data in collaboration with the research structures and the graduate schools. **ORIGINALITY** There is a growing body of research papers on data management and curation. Produced along with PhD dissertations, little is known about the characteristics of this material, in particular in social sciences and humanities and the impact on the role of academic libraries.

Received: 02/27/2015 Accepted: 06/09/2015

Correspondence: Joachim Schöpfel, GERiiCO Laboratory, University of Lille 3, 3 Rue du Barreau, 59650 Villeneuve-d'Ascq, France, joachim.schopfel@univ-lille3.fr

IMPLICATIONS FOR PRACTICE

An improved knowledge on research data submitted with PhD dissertations can help to:

1. Assist and advise PhD students on how to manage their research data (data management plan).
2. Assist PhD students in depositing their research data.
3. Prepare and select research data for long-term preservation.
4. Curate research data in open repositories.
5. Design data repositories and related content mining tools.

INTRODUCTION

Open access to PhD dissertations[1] is on the agenda of academic libraries. There are some good reasons to make PhD dissertations widely available to the scientific community as well as to the general public. One reason is that many PhD projects are publicly-funded research and as such should be made accessible to society. Another reason is that their free and large dissemination will increase the visibility and impact of the institution's scientific output. Other arguments in favour of open access are their quality (validation by jury), their representativeness (coverage of all disciplines), their novelty and originality, and their rich content, including exhaustive and detailed literature reviews and bibliographies.

Today, the rapid development of data-driven research (e-Science) and the debate on open data and re-use of research results has led us to discover another challenge in the field of PhD dissertations, beyond the debate on open access and embargo, i.e. the existence of large amounts of small data produced by the PhD candidate and partly submitted together with the text of the dissertation. These small data are the topic of our paper. We wonder how these data can be made available in the context of open access and open data policies, what are the potential barriers, and how academic libraries could contribute to this challenge.

Our exploratory study is part of a digital humanities research project on ETDs and research data, with two objectives: (a) create a campus-based service together with the academic library, the graduate school and research laboratories, to assist PhD students in research data management (RDM), preservation and dissemination of their research results; and (b) develop content mining tools for the further exploitation of these data.

[1] In the following we shall use the term "PhD dissertation" to designate the document submitted in support of candidature for the academic degree of doctorate, as synonym with "PhD or doctoral thesis".

LITERATURE REVIEW

Research results produced by PhD students could contribute to e-Science, i.e. cyberinfrastructures enabling data-intensive scientific discovery (Hey, Tansley, & Tolle, 2009). The PhD dissertations could become "windows for the scientist" not only to understand but also to reproduce and extend scientific results (Lynch, 2009), in so far as they could integrate research data that could be enriched, updated, extracted, shared, aggregated and manipulated (McMahon, 2010). They could become live documents.

They could—but there are two barriers. First, to become live documents and "windows for the scientist," dissertations must be freely available in open access, deposited in institutional or other repositories, and disseminated with sufficient user rights to allow re-use. However, up to now a significant portion of the digital dissertations are not online, not open, not freely available but embargoed or under restricted access (Schöpfel et al., 2015). Today, more than half of all open repositories contain theses and dissertations,[2] and the technical and political environment globally supports open access to academic works but universities, graduate schools, and academic libraries still have a long way to go to promote and ensure open access to all PhD dissertations.

The second barrier is the fact that research data related to PhD dissertations are largely "dark data," i.e. "data that is not easily found by potential users (…) unpublished data (and) research findings and raw data that lie behind published works which are also difficult or impossible to access as time progresses" (Heidorn, 2008, pp.281 and 285). The question of research data produced along with PhD dissertation has not attracted much attention so far (see the overview by Schöpfel et al., 2014). Often, studies on data related to electronic theses and dissertations, like Song (2007), do not distinguish between text and research data and thus miss the point that these documents bear hidden treasures waiting to be discovered, disseminated, and re-used. Perhaps the recent IMLS funded project ETDplus[3] at the Educopia Institute at Atlanta, Georgia, will provide more insight into this question.

These data, defined as re-usable research results, collected, observed, or created for purposes of analysis to produce original research results, are part of what is called "small data," produced in a large variety of formats, sources, and types. Research results may be presented as tables, graphs, etc. in the paper or as additional material (appendix). However, there is no

[2] See the international Directory of Open Access Repositories OpenDOAR http://www.opendoar.org/

[3] http://educopia.org/research/grants/etdplus

clear or unique definition of the term "research data." Following the OMB Circular 110,[4] research data can be considered as "the recorded factual material commonly accepted in the scientific community as necessary to validate research findings." The international directory for research data repositories *re3data*[5] distinguishes between fourteen different types of data (archived data, audio-visual data, configuration data, databases, images, network-based data, plain text, raw data, scientific and statistical data formats, software applications, source code, standard office documents, structured graphics, and structured text) but admits that there are other categories in the more than 1,200 indexed repositories.

In the past, print theses and dissertations have regularly been submitted together with supplementary material and data, in various formats and on different supports (print appendices, punched cards, floppy disks, audiotapes, slides, CD-ROMs). In the new ETD infrastructures, such material is sometimes but not always submitted and processed together with the text files or as supplementary files in different formats, depending on disciplines, research fields, and methods. The ETDplus project states that these "complex digital objects (e.g., software, multimedia files, digital art, and other material that sometimes is integral to the thesis or dissertation itself […])" are often not collected or preserved.[6]

When this material is submitted as a kind of data appendix, the dissertation becomes a "data vehicle," where data are published together with the dissertation or as a part of it. Sometimes the data are available on a distant server and without the text of the dissertation, transforming the dissertation in a "gateway to data." Yet, too often the data are simply not available, or data, methodology, tools, primary sources are mingled, not indexed, badly described, unrelated with the text, unconnected with other files.

Obviously, these "supplemental research data and complex digital objects" need curation and management to remain accessible and interpretable over time. The data management includes metadata and long-term preservation (Neuroth, Strahmann, Oßwald, & Ludwig, 2013). For young scientists and PhD students, learning how to design and implement a data management plan (RDM) is even more important in so far as more and more funding bodies (such as the European Commission or the French National Research Agency) evaluate the existence and quality of RDMs in research project proposals.

[4] https://www.whitehouse.gov/omb/circulars_a110#36

[5] http://www.re3data.org/

[6] http://educopia.org/sites/educopia.org/files/grantdocuments/ETDplus_Narrative.pdf

Description and preservation of digital objects are part of the work of traditional academic libraries. For this reason, they generally consider research data curation and management as a new challenge, a kind of new frontier for the development of their campus services, either on a local level or as part of a scientific network (CLIR 2013). Thus, they are engaged in the development of data repositories (Lynch, 2014; Newton, Miller, & Bracke, 2011), in campus-wide surveys (Simukovic, Kindling, & Schirmbacher, 2014) or in research projects like the *Policy RECommendations for Open Access to Research Data in Europe* (RECODE)[7] project or the global registry of research data repositories covering research data repositories from different academic disciplines (re3data.org).[8]

METHODS

The survey was conducted at the University of Lille 3, a large social sciences and humanities campus in the Northern part of France, with 19,000 students and nearly 500 PhD candidates in three graduate schools and (in 2013) 55 doctoral degrees.[9] The survey was conducted between November 2014 and January 2015. The results were presented and discussed during a research seminar at the Lille European Social Sciences and Humanities Research Institute[10] in February 2015.[11]

Our sample contains all digital dissertations from the University of Lille 3 available through the national dissertation portal[12] and the Lille 3 institutional repository.[13] We completed this list with older print dissertations, especially in the Lille 3 fields of excellence (History, Archaeology, Egyptology, Linguistics, Psychology) to obtain a sample of nearly 300 dissertations (Figure 1, following page).

The sample consisted of 88 digital dissertations (31%) and 195 print dissertations (69%), from 1987 to 2013. Altogether, these 283 dissertations represent 30% of all dissertations from

[7] http://recodeproject.eu/

[8] http://www.re3data.org/

[9] For more information about the French PhD dissertation infrastructure, see Paillassard, Schöpfel, and Stock (2007) and Giloux & Mauger-Perez (2008).

[10] http://www.meshs.fr

[11] http://drtdshs2015.sciencesconf.org/resource/page/id/1

[12] http://www.theses.fr

[13] http://hal.univ-lille3.fr/

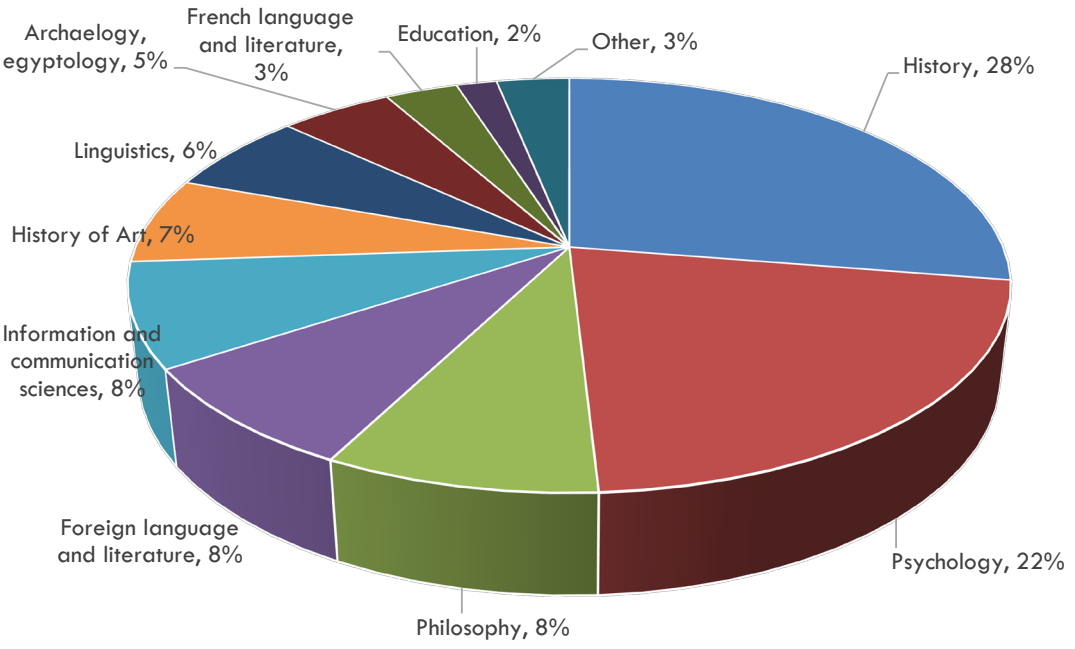

Figure 1. Scientific disciplines of the survey sample (N=283 dissertations)

1987 to 2013 from the University of Lille 3. In our sample, History, Psychology, Philosophy, Foreign Languages and Literature (English and American, Spanish, Slavonic, Hebrew…), Information and Communication Sciences (including Library Sciences), History of Art, Linguistics, Archaeology, and Egyptology were the most represented disciplines, followed by French Language and Literature, Education, Sociology, Musicology, and others.

All dissertations have been analysed either in digital or print format or on microform. Each dissertation has been checked by at least one of the authors, either in the library holdings (print or microform) or on the institutional repository server. We tried in particular to identify research data added to the end of the dissertation.[14] The evaluation checklist was based on the Berlin survey on research data (Simukovic et al., 2014) and included data sources, data types, size (or volume) of data, and the mode of publishing (appendix,

[14] In this paper we shall use "appendix" as a generic term for document(s) attached to the end of a dissertation, without distinguishing between "annex" and "appendix" as in the French usage.

separate volume). Our results will be archived as a spreadsheet for further re-use.[15] As our study has an exploratory character, we adopted a large and pragmatic definition of research data, compliant with the widely accepted OMB and re3data definitions cited above or the Educopia concept of "research data and complex digital materials." This may not be completely satisfying but allows for a realistic study of submitted materials and data.

RESULTS

In our sample of 283 dissertations, 188 contain one or more appendices with some kind of research data (66%). The length of these appendices varies widely, from 5 to 829 pages, with a median of 81 pages, and totalling more than 25,000 pages (Figure 2).

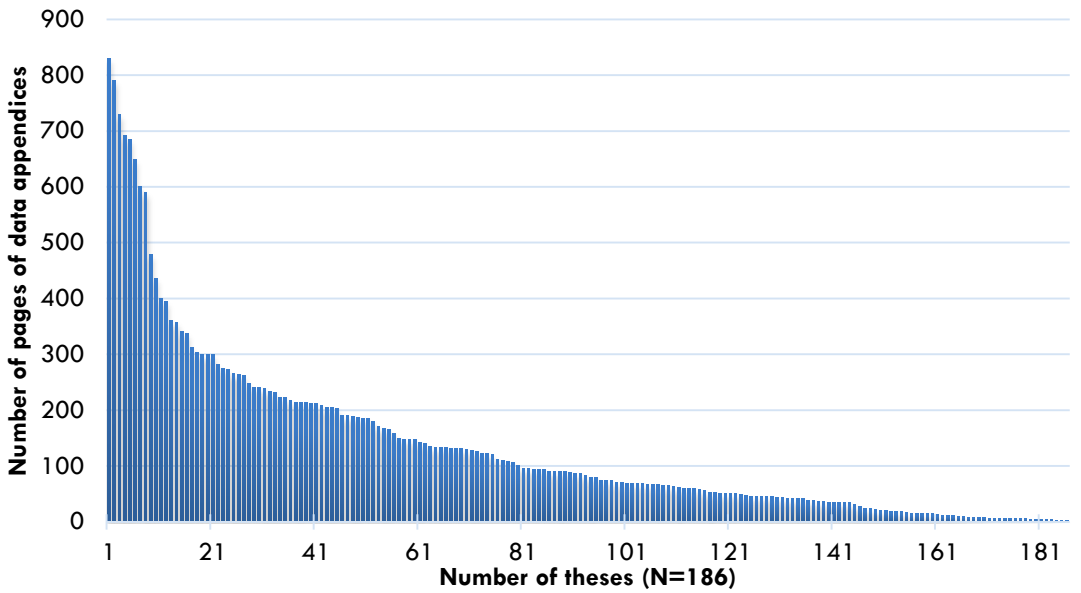

Figure 2. Size of data appendices (in pages, N=188 dissertations)

[15] Forthcoming as part of the thematic collection "GreyNet Enhanced Publications Project" in the Dutch EASY repository at https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:51624

Even if each appendix holds some kind of research data, this does not mean that one can find research results (data) *stricto senso* on all pages. Some pages contain empty questionnaires or survey forms, experimental procedures, bibliographies etc. which cannot be considered as data.

Disciplines

The distribution of disciplines per dissertation with appendices is more or less the same than for the overall sample (see Figure 3). The linear determination coefficient between both variables is high (R^2=.91).

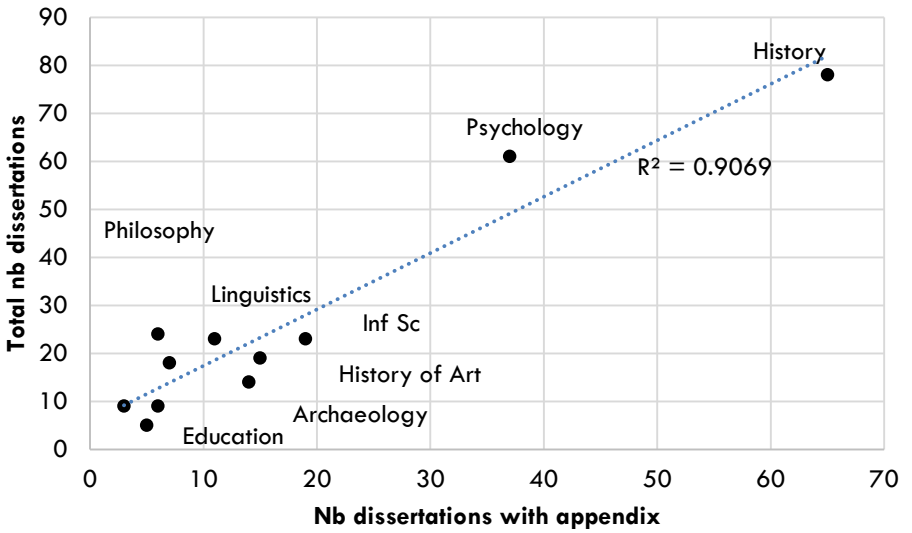

Figure 3. Scientific disciplines of dissertations (overall sample; with appendix)

The differences are not significant—in some domains such as Psychology, Philosophy and Linguistics, there are fewer dissertations with data appendices than the average (66%); in others there are relatively more (Information Sciences, History of Art). In Education and in Archaeology and Egyptology, all dissertations of our sample contain some form of data appendices.

The differences between disciplines are elsewhere (Figure 4, following page). Some disciplines produce rather large appendices, with an average number of pages above the mean of the whole sample, such as History of Art, Education, Archaeology and Egyptology,

while others most often contain shorter appendices (Psychology, Philosophy, Language and Literature, etc.).

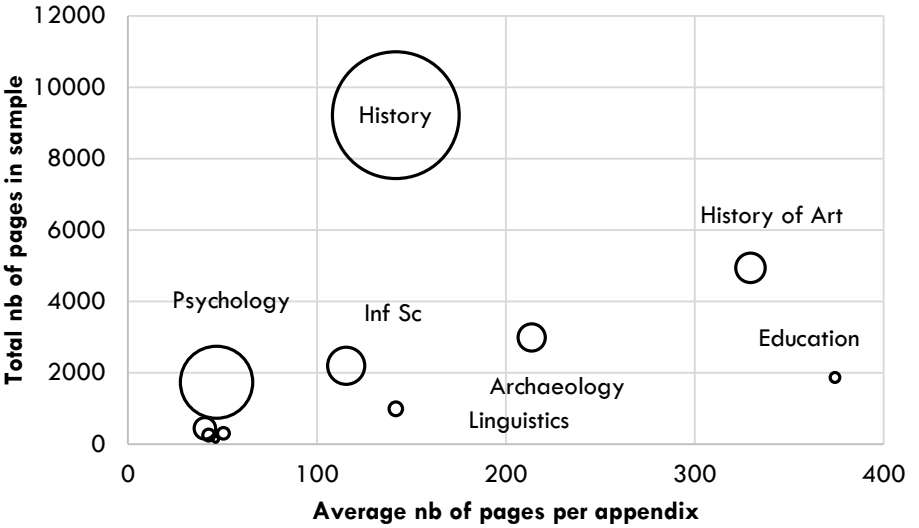

Figure 4. Size of appendices (circle size = number of dissertations, N=188)

Support, presentation and format

The French official guidelines for PhD dissertations[16] do not specify how to structure or present an appendix. Also, some dissertations have poor or no table of contents for their appendices, like a recent dissertation in History[17] with a table of content for the text volume but not for the two volumes that contain rich material, including 1,581 figures and images. As mentioned above, one third of our samples are electronic dissertations. Compared to the print dissertations, they often contain slightly more appendices (Figure 5, followin page).

Also, electronic dissertations often do not separate text and appendices but glue them together into the same file (52%). For comparison: 30% of the print dissertations clearly separate text and appendices in different volumes. This difference is highly significant (χ^2=14.32).

[16] http://www.abes.fr/Media/Fichiers/Theses-Fichiers/theses.fr/Guide-du-doctorant-2013-pdf

[17] De Salas, A. (2010). *L'iconographie de sainte Anne en Espagne à la fin du Moyen Age*. Université de Lille 3

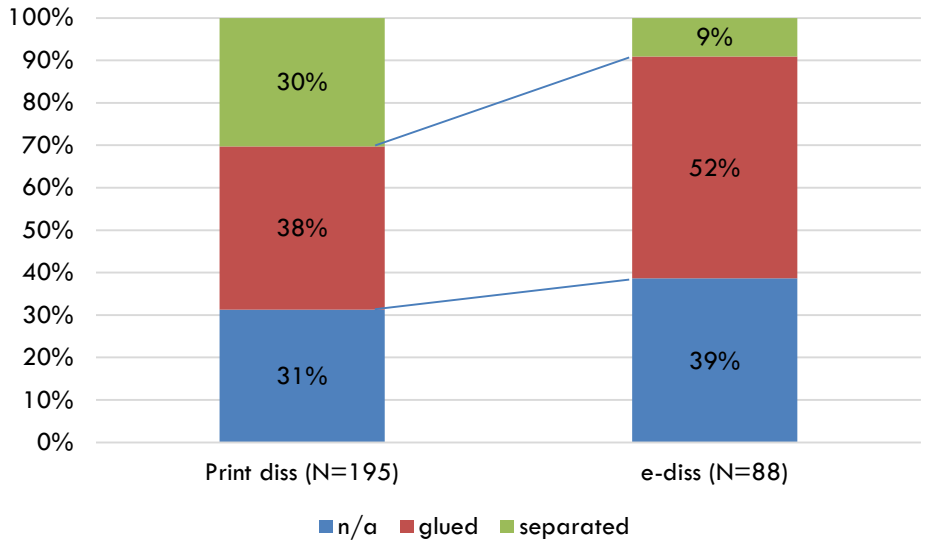

Figure 5. Link between text and appendices (in %, N=283)

All files of digital PhD dissertations must be deposited with the text, and the French national computer centre for Higher Education[18] maintains a list of accepted file formats for long term preservation.[19] However, nearly all files are in PDF (image or text), and other formats are very rare. In our sample, only one dissertation has been submitted with video and audio files on CD-ROM.[20]

Some dissertations demonstrate a real effort of data management and curation by the PhD student. For instance, a dissertation in Egyptology on late Egyptian steles[21] contains an exhaustive inventory of those steles on CD-ROM with indexing of geographical origin, general characteristics, specific particularities and dating, together with the transcription of the inscription and a justificatory supporting the provenance of the stele. The PhD student also delivers a user manual for the navigation in the database.

[18] CINES https://www.cines.fr

[19] https://www.cines.fr/archivage/des-expertises/expertise-formats/liste-des-formats-archivables/

[20] Barth, C. (2010). *Des TIC comme vecteur matériel et symbolique de rationalisation et modélisation de la vie domestique: le cas de « l'intelligence ambiante »*. Université de Lille 3.

[21] De Visscher, C. (2011). *Étude typologique des stèles égyptiennes tardives : essai d'analyse qualitative et quantitative*. Université de Lille 3.

Another dissertation in Linguistics[22] presents a diachronic analysis of the vocabulary from 49 political speeches and 10 manifestos, pamphlets and articles, with lexical analyser software (*Wmatrix* corpus analysis and comparison tool). The appendix contains the complete list of all words with their frequency of usage ranking.

Dissertations in History, especially for studies on historical social groups, sometimes contain detailed and well-structured biographical information presented like a database. One example for this "prosopographical" approach: a dissertation on the Renaissance elite of the old Flemish town of Douai[23] with biographical records of 423 aldermen, with structured information about, among others, place and date of birth, date of death, mandate period, noble titles, and occupation.

Research data sources

The PhD students used a great variety of sources for their scientific work. Based on the Simukovic et al. (2014) survey, we identified three major data sources, i.e. archives, surveys, and text samples or corpora (Figure 6).

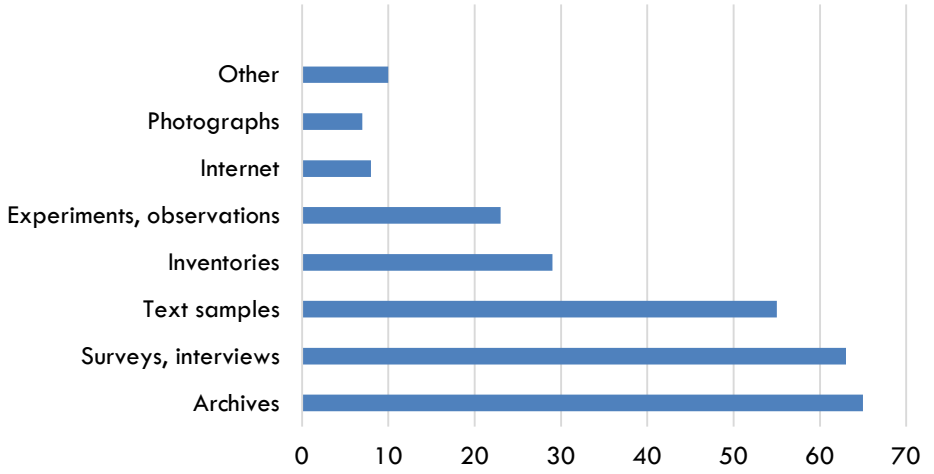

Figure 6. Data sources per dissertations (N=188)

[22] L'Hôte, E. (2011). *The language of politics: a corpus-based cognitive analysis on new Labour discourse (1994-2007)*. Université de Lille 3.

[23] Duquenne, F. (2011). *Un tout petit monde : les notables de la ville de Douai du règne de Philippe II à la conquête française (milieu du XVIe siecle-1667) : pouvoir, réseaux et reproduction sociale.* Université de Lille 3.

Other less exploited sources are inventories, experiments and observations, Internet and photographs. The distribution of data sources is to some extent specific for each discipline. Here are some examples of heavily used sources:

- History: archives, text samples
- Psychology: surveys, experiments
- Philosophy: text samples
- Foreign Languages and Literature: text samples
- Information and Communication Sciences: surveys, text samples, Internet
- History of Art: inventories
- Linguistics: text samples, surveys
- Archaeology and Egyptology: inventories, photographs

These are typical research data sources for the social sciences and humanities. Compared to the Berlin survey, other data sources like observations, simulations, statistics, reference data, or log files (usage data) are unusual or missing.

Typology of research data

Finally, which are the research data present in the appendices? Our evaluation reveals several different and heterogeneous data types (Figure 7).

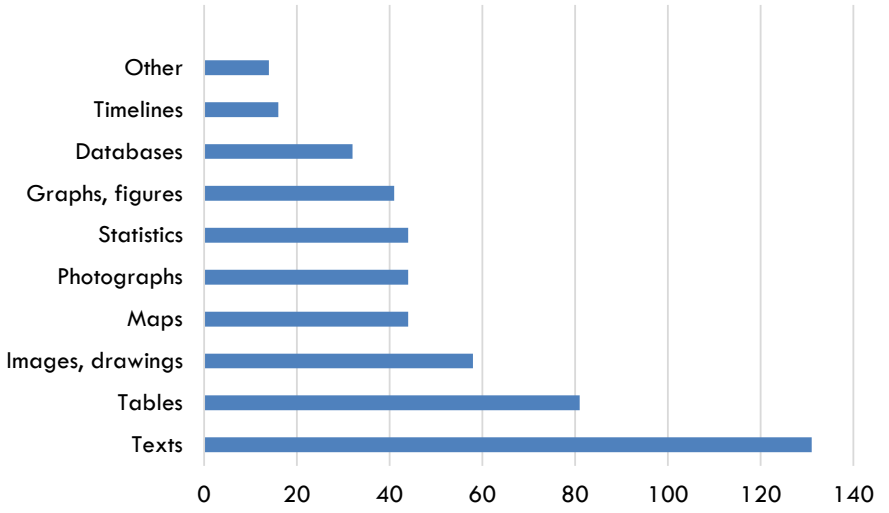

Figure 7. Data types, per dissertations (N=188)

Again, text samples are the most important data type, followed by tables (spreadsheets), images (including drawings and posters), maps, photographs, statistics, graphs (including figures, charts and visualisations), databases, and timelines (chronologies). We only found one dissertation with audio-visual media (interviews),[24] and we have not found any dissertation with geolocation data. Also, there are some discipline-specific data type profiles.

In some disciplines, one or two data types are predominant. This is the case in Philosophy, Linguistics, and Foreign Languages and Literature, where text samples represent more than half of the data. Other disciplines are characterized by a wide number of different research data. Some examples:

- History: ten different data types, including text (21%), tables (15%), and images (14%).
- Information and Communication Sciences: ten different data types, including text (33%), tables (17%), and graphs (13%).
- Psychology: nine different data types, including tables (29%), and statistics (28%).
- Archaeology and Egyptology: nine different data types, including photographs (21%), maps (17%), and images (17%).
- History of Art: eight different data types, including text (29%), and images (18%).

The research data are very different. Some examples: a great number of images and photographs on the religious life in the French town of Etaples from the beginnings to 2000,[25] statistics on prisons and prisoners in Northern France during the French Third Republic,[26] the mapped tours and comments of children in a dissertation on two exhibitions,[27] or a large corpus of old documents and archaeologic findings for the reconstruction of the organisation of banquets in Anglo-Saxon England.[28]

[24] A dissertation in information sciences about new information and communication technologies in domestic life: Barth, I. (2010). *Des TIC comme vecteur matériel et symbolique de rationalisation et modélisation de la vie domestique*. Université de Lille 3.

[25] Baudelicque, P. (2000). *Histoire religieuse d'Etaples des origines à l'an 2000*. Université de Lille 3.

[26] Lambin, S. (2013). *Les prisons et prisonniers dans le Nord sous la IIIe République*. Université de Lille 3.

[27] Martin, T. (2011). *L'expérience de visite des enfants en musées de sciences dans le cadre des loisirs : logiques d'interprétation et enjeux d'un dispositif communicationnel.* Université de Lille 3.

[28] Gautier, A. (2004). *« Pr weras drincath », la ou les hommes boivent : le festin dans l'Angleterre anglo-saxonne, V-XIe siecles.* Université de Lille 3.

Some data types are present in all disciplines, like text samples, images, tables or graphs, and figures. Others, in particular inventories or audio-visual material, are at least in our sample specific for one or two disciplines. We compared print dissertations and e-dissertations and performed a chi-squared test but found no significant differences neither for research data sources nor for data types (on 0.05 level). Obviously, these differences are more related to disciplinary methodologies than to support.

DISCUSSION

Shortfalls

Even if the sample comprises nearly 300 dissertations from more than 50 disciplines and sub-disciplines, it does not pretend to be representative. First of all, as this study is part of a research project in digital humanities, dissertations from sciences, technology, or medicine were excluded. Secondly, some domains of social sciences and humanities are under-represented or missing, like Sociology, Economics, Cultural Studies, Law, and Politics. This study has an exploratory character with the intention to produce illustrating examples and figures for a better understanding of the challenge.

Another methodological problem is the difficult distinction between data sources and data types and, more generally, the definition of what research data is and what it is not. As we told above, we adopted a more pragmatic approach, admitting that the concept of research data depends on scientific fields and methods, is more or less open, and not clearly discernible from data sources. For instance, are photos of archaeological inventories primary sources students used for their analysis, or results of their research, or both? Our approach was to identify and describe types of research data that are potentially re-usable which means that they may become, together with the dissertation, sources of further research and future research data.

A last shortfall is the limitation to appendices. Many research data are available in appendices. However, sometimes significant data amounts are part of the text and not clearly distinguished. For instance, a dissertation in History on the dairy economy in North France[29] lists 13 maps, 54 promotional documents, 124 drawings, 52 photographs, 21 graphs, and 4 spreadsheets, all distributed throughout the whole text. In some dissertations in History of Art, the artworks are documented in a special volume but not marked as an appendix.

[29] Delbaere, N. (2007). *L'économie laitière dans le Nord-Pas-de-Calais, de l'âge rural à l'âge des marques*. Université de Lille 3.

Barriers to open data

The research data we identified in our sample of dissertations could be re-used for further research in a different way. Text samples, for example, could be objects for lexical analyses or text mining or even for specific historical investigations, like prosopography. Photographs, images, or drawings could be organized as image databases and/or online inventories. Tables, statistics, and already well organized databases could be useful for data mining or aggregated for meta-analyses.

However, this potential re-use requires data management and curation. Our study reveals three barriers to open data:

- *Incomplete, inadequate or missing description of the whole datasets and/or individual data.* In some dissertations, especially in History, History of Art and Archaeology, inventories, photographs, maps, etc. are well described and indexed. But these are exceptions and often descriptions are simply missing.
- *Missing organisation.* Research data are presented without any structure or organisation, often together with other, not re-usable material in a kind of information mash-up not suitable for further research.
- *Inadequate format.* In print copies, this means that data are not clearly separated from the dissertation text. In electronic dissertations, this means that data and text are glued together in a PDF file instead of being separated and published in adequate file formats (spreadsheets, image files, text files, database formats, XML…).

Other problems are related to the choice of media, e.g. compact disc, DVD, online server, USB flash drive, etc. For instance, the dissertation on Egyptian steles informs readers about an online database with restricted access but does not provide the login and password. For some retro-digitized dissertations, the online version does not include the data appendix submitted together with the print version.

Our empirical data don't tell us if the PhD students conducted a data management plan. Yet, other survey results[30] reveal that probably most of them didn't so far, even if they are interested in data management and to some extent in support of sharing at least some data on Internet.

[30] Simukovic et al. (2014) and two recent, unpublished surveys from the French Universities of Strasbourg and Lille 3 (forthcoming).

Legal aspects

Our study team did not include a legal expert. Yet, a superficial evaluation of legal aspects reveals at least two legal problems related to the deposit and dissemination of research results in dissertations:

- *Privacy issues.* Some appendices contain personal data about living or dead people, historical persons, or unknown (anonymous) people. These may be survey data, experiments, interviews, biographies, etc. In so far as the information allows identifying individual persons, at least with regards to the French law, it needs special processing and careful handling.
- *Third party copyright.* Some dissertations contain material that is protected by copyright and cannot be reproduced or disseminated without authorization, even by fair use or copyright exceptions (short citation, research…). This may be text samples, maps, photographs, copies from books, etc. (material not created by the PhD student him/herself).

Sometimes, the authorship of the research data remains uncertain.

CONCLUSION

Our exploratory survey on research data in PhD dissertations in social sciences and humanities reveals a large variety and great richness of data sources and types submitted as appendix with the text of the dissertations. Many, if not all, of these data could be of real value for further research. These data could be used to create image databases, digital maps collections, or digital libraries with manuscripts, archival material, and other text samples open for text mining tools. Results from experiments and surveys could be published in a way that allows for re-use, data mining, and automatic meta-analysis on different datasets. Research results could thus become new data sources and generate further research.

However, re-use and content mining requires management and curation of the research data, and our sample leads us to believe that advice and assistance will be necessary for PhD students to prepare their data in an adequate way. Adequate means at least:

- *Clear separation of text and data.* Digital research data must be submitted in different and separate files.
- *Structuration of the research data*, with detailed and organized tagging (markup) of the datasets.

- *Metadata of good quality*. The data must be described in a standard language and format, with sufficient detail for retrieval and data mining.
- *Deposit in original format.* Data should be submitted in their original and if possible, open format (and not in PDF), to facilitate long-term preservation and re-use.

These are general rules. However, the empirical evidence of this study suggests that assistance and advice for PhD students to help them manage their research data must go beyond general rules and recommendations. Not all doctoral projects produce research data. Not all data are submitted with the dissertation to back up the research in the dissertation or to further explain and clarify the matter. Not all data can be re-used especially, but not only, for legal reasons. And finally, even if our sample is not representative, it seems obvious that many characteristics of data sources and types have strong relationships with disciplinary methods, topics, and approaches.

Further investigation is needed, in particular with regards to disciplines like sociology, economics, cultural studies, law, and politics, also on formats, metadata, and legal aspects. Yet, the results of our study suggest that in any case, a campus-based data management service for PhD students should be aware of strong disciplinary particularities and must be as close as possible to their research practices and needs. For this reason, the best solution would be a service based on a partnership between academic libraries, graduate schools, and research laboratories. We will continue our analysis on research data and digital objects in dissertations, with a larger sample of digital dissertations and more disciplines, and present the results during an international conference on grey literature at the Royal Dutch Academy of Arts and Sciences in December 2015 (Schöpfel, et al., forthcoming); meanwhile, we will prepare a white paper on research data in dissertations that will contribute to a three-tiered service on our campus, including an integrated service of education, consultation, and infrastructure,[31] with three guiding principles: a strong commitment to sharing and open access, a disciplinary approach as one size does not fit all, and a close collaboration between the academic library (future learning center), the graduate schools, and the research laboratories.

[31] Comparable to the approach of Reznik-Zellen, Adamick, and McGinty (2012).

REFERENCES

CLIR (2013). *Research data management: Principles, practices, and prospects*. Report, Council on Library and Information Resources, Washington D.C. Retrieved from http://www.clir.org/pubs/reports/pub160

Giloux, M., & Mauger-Perez, I. (2008). A new French circuit for the electronic theses. In *ETD 2008 11th International Symposium on Electronic Theses and Dissertations, 4 - 7 June 2008*, The Robert Gordon University, Aberdeen, UK. Retrieved from http://www4.rgu.ac.uk/etd/programme/page.cfm?pge=45696

Heidorn, P. B. (2008). Shedding light on the dark data in the long tail of science. *Library Trends*, *57*(2), 280-299. Retrieved from https://www.ideals.illinois.edu/bitstream/handle/2142/10672/heidorn.pdf?sequence=2

Hey, T., Tansley, S., & Tolle, K. (Eds.) (2009). *The fourth paradigm. Data-intensive scientific discovery*. Redmond, WA: Microsoft Corporation. Retrieved from http://research.microsoft.com/en-us/collaboration/fourthparadigm/

Lynch, C. (2009). Jim Gray's fourth paradigm and the construction of the scientific record. In T. Hey, S. Tansley, & K. Tolle (Eds.), *The fourth paradigm. Data-intensive scientific discovery* (pp. 177-183). Retrieved from http://research.microsoft.com/en-us/collaboration/fourthparadigm/

Lynch, C. (2014). The need for research data inventories and the vision for SHARE. *Information Standards Quarterly*, *26*(2), 29+. http://dx.doi.org/10.3789/isqv26no2.2014.05

McMahon, B. (2010). Interactive publications and the record of science. *Information Services and Use*, *30*(1), 1-16. Retrieved from http://iospress.metapress.com/content/f4th457822023783/fulltext.pdf

Neuroth, H., Strahmann, S., Oßwald, A., & Ludwig, J. (Eds.) (2013). *Digital curation of research data. Experiences of a baseline study in Germany*. Glückstadt: vwh. Retrieved from http://www.nestor.sub.uni-goettingen.de/bestandsaufnahme/Digital_Curation.pdf

Newton, M. P., Miller, C. C., & Bracke, M. S. (2011). Librarian roles in institutional repository data set collecting: Outcomes of a research library task force. *Collection Management, 36*(1), 53-67. http://dx.doi.org/10.1080/01462679.2011.530546

Paillassard, P., Schöpfel, J., & Stock, C. (2007). Dissemination and preservation of French print and electronic theses. *The Grey Journal*, *3*(2), 77-93. Retrieved from http://archivesic.ccsd.cnrs.fr/sic_00380488/en/

Reznik-Zellen, R., Adamick, J., & McGinty, S. (2012). Tiers of research data support services. *Journal of eScience Librarianship*, *1*(1), 27-35. http://dx.doi.org/10.7191/jeslib.2012.1002

Schöpfel, J., Chaudiron, S., Jacquemin, B., Prost, H., Severo, M., & Thiault, F. (2014). Open access to research data in electronic theses and dissertations: An overview. *Library Hi Tech*, *32*(4), 612-627. http://dx.doi.org/10.1108/LHT-06-2014-0058

Schöpfel, J., Prost, H., Piotrowski, M., Hilf, E. R., Severiens, T., & Grabbe, P. (2015). A French-German survey of electronic theses and dissertations: Access and restrictions. *D-Lib Magazine*, *21*(3/4). http://dx.doi.org/10.1045/march2015-schopfel

Schöpfel, J., Juznic, P., Prost, H., Malleret, C., Cesarek, A., & Koler-Povh, T. (2015 forthcoming). Dissertations and data (keynote address). In *GL17 International Conference on Grey Literature*, 1-2 December 2015, Amsterdam.

Simukovic, E., Kindling, M., & Schirmbacher, P. (2014). Unveiling research data stocks: A case of Humboldt-Universität zu Berlin. In *iConference, 4-7 March 2014, Berlin*, (pp. 742-748). Retrieved from http://hdl.handle.net/2142/47259

Song, I. (2007). Promoting open access to scholarly data: A case study of the electronic thesis and dissertation (ETD) project at the Simon Fraser University Library. *Data Science Journal*, *6*(suppl.), S70-S78. http://dx.doi.org/10.2481/dsj.6.S70

All websites were accessed in February 2015.